WAVELET ANALYSIS OF IRREGULARLY SPACED DATA AND ITS SPATIO-TEMPORAL EXTENSION

著者	Matsuda Yasumasa
journal or publication title	DSSR Discussion Papers
number	30
page range	1-21
year	2014-11
URL	http://hdl.handle.net/10097/65006

DSSR

Discussion Paper No. 30

WAVELET ANALYSIS OF IRREGULARLY SPACED DATA AND ITS SPATIO-TEMPORAL EXTENSION

YASUMASA MATSUDA

November, 2014

Data Science and Service Research Discussion Paper

Center for Data Science and Service Research
Graduate School of Economic and Management
Tohoku University
27-1 Kawauchi, Aobaku
Sendai 980-8576, JAPAN

WAVELET ANALYSIS OF IRREGULARLY SPACED DATA AND ITS SPATIO-TEMPORAL EXTENSION

YASUMASA MATSUDA

Abstract. This paper proposes an orthogonal basis expansion by Haar wavelets for irregularly spaced spatial data and conducts its spatio-temporal extension when spatial points are not necessarily the same at each temporal point. One crucial contribution of this paper is discovery of algorithm to construct orthogonal basis for any irregularly spaced data via 2 dimensional Haar wavelets. We construct a spatial model by imposing a prior distribution on coefficients of the orthogonal expansion. In conducting spatio-temporal extension, we fit a simple autoregressive model not to data itself but to the coefficients of the orthogonal expansion, which makes it possible to handle spatial sampling points that are not necessarily the same at each temporal point. The spatio-temporal model can express nonstationary spatial behaviors by allowing parameters dependent spatially. Advantage of the models is superb parameter estimation efficiency that is attained by independency of wavelet transform providing efficient Gaussian likelihood evaluation. Finally we demonstrate the wavelet modeling for land price data in Tokyo from 2003 till 2014, yearly series of public land prices per m^2 over about 8000 sampling points scattered irregularly all over Tokyo areas.

1. Introduction

Spatial or spatio-temporal data have recently been attracting growing attentions in areas of both natural and social sciences by rapid progress of data collection technologies. This paper considers a modeling for spatial data observed at sampling points irregularly scattered over so broad region that stationary covariances cannot be assumed, and aims to conduct its spatio-temporal extension.

Classical geostatistical methods provides several kinds of spatial covariance models such as Spherical, Exponential, Gaussian, and Matérn classes to analyze spatial data (Zimmerman and Stein, 2010). Estimation of parameters in the covariance models by Gaussian likelihood requires evaluation of determinant and inverse of the covariance matrix whose dimension is as many as the sample size. It follows that exact Gaussian likelihood is in practice difficult to evaluate especially for large data set.

In order to avoid the computational difficulties, several methods for approximations have been proposed in the literatures. Typical ones that represent them are the covariance tapering by Kaufman et. al. (2008), the predictive process approach by Banerjee et. al. (2008) and the composite function approach by Bai et. al. (2012). The three methods commonly employ likelihood functions approximated in clever ways to avoid the difficulties caused by large dimensional covariance matrices. Although they work reasonably well for large spatial data set, trivial extensions to

Key words and phrases. forecasting. Haar wavelet. Kalman filter. kriging. state space model. Whittle likelihood function.

spatio-temporal data are not efficient enough to approximate the likelihoods for as many as several ten thousand samples, which is normally supposed in spatio-temporal extensions.

Reduced-rank random effects approach, a method of dimension reduction to let Gaussian likelihood evaluation more efficient, has been used for analysis of large spatial data set in the literatures as an alternative (Wikle, 2010). Let Z be a n-dimensional observation vector possibly for large n. For a $n \times p$ regressor matrix H with $p < n$, the approach employs the following multiple regression models

$$Z = H\beta + \varepsilon,$$

where β is assumed to follow normal distribution with mean 0 and variance Σ, which is possible to regard as a prior distribution for β. There are variety of choices for H and Σ, which specify practical performances of modeling for large spatial data set.

Let us introduce principal three approaches for the choice of H and Σ in spatial data modeling. The first one is Karhunen-Loéve expansion approach. Assuming that $\varepsilon = 0$, we have

$$Var(Z) = H\Sigma H',$$

which means that H and Σ are obtained by a singular value decomposition for $Var(Z)$ as orthogonal and diagonal matrices, respectively. Though the approach gives a natural interpretation for the selection of H and Σ in the same way as principal component analysis does, we have to tackle the following non-trivial problems in practical applications. $Var(Z)$ is in practice unknown and should be estimated. The estimation is not trivial except for the case such as independent realizations of spatial data is available. The matrix H constructed by the eigenfunctions are defined just over sampled points and requires some technique to apply it for kriging. For papers handling those in Karhunen-Loéve expansion approach, see, e.g., Chen and Huang (2013) and references therein.

The second one is non-orthogonal basis function approach that specifies the mean function $H(s)\beta$ at a spatial point s with

$$\sum_{j=1}^{p} h(s, r_j)\beta_j,$$

for some kernel function h. Since the Gaussian likelihood is calculated with a $p \times p$ matrix operation in the model (see, e.g. Wikle, 2010), the computational burden in the approach data will be greatly reduced, if $p << n$ is chosen. As long as $p < n$ is designed to be moderate, the model can express flexible class of nonstationary spatial covariance structures with efficient estimation performances. Adequate choice of kernel and Σ, which is difficult in practice, would provide good modeling with efficient performances. Papers such as Shi and Cressie (2007), Cressie and Johannesson (2008) are typical ones categorized in this approach employing W-wavelets and local bisquare function as a basis matrix H, respectively. Stroud et. al. (2007) attempted spatio-temporal modeling in this approach.

The final one is orthogonal basis expansion approach composed of orthogonal vectors $H = (h_1, \ldots, h_p)$ that satisfies $h_i' h_j = \delta_{ij}$. The orthogonal basis expansion for spatial data would provide Gaussian likelihood evaluation that requires one dimensional operation, if we assume Σ is diagonal. As a result, p may be chosen as many as sample sizes with sufficient approximation accuracy for the mean function term. The point in this approach is how we construct orthogonal basis for irregularly spaced data. The traditional orthogonal basis such as Fourier, wavelets, Hermite polynomials and so on can work only when sampling points are regularly spaced to constitute mesh data. The orthogonal property of them is usually destructed under irregular sampling. Paciorek (2007) employed Fourier basis as an orthogonal basis by binning data locations, the modification of data locations to let them in grid locations. One principal theme in this paper is how we find orthogonal basis under irregular sampling without binning.

This paper adopts the final approach for choice of H. Specifically we aim to construct block-orthogonal basis H such that $H'H$ is block diagonal by modifying Haar wavelet basis under irregular sampling. Moreover we aim to conduct spatio-temporal extension of the Haar wavelet model from spatial data to spatio-temporal data. In conducting the extension, we do not assume that sampling spatial points are the same for each temporal point.

The contributions in this paper are summarized in the following three points. The first one is construction of block-orthogonal basis matrix H for irregularly spaced data, which is obtained by modifying Haar wavelets. The second one is discovering efficient way of estimation of parameters by wavelet transform with consistency proof under fixed domain asymptotics. The use of block-orthogonal basis expansion plays a key role in letting Gaussian likelihood evaluation be an efficient form requiring at most three dimensional matrix operations. Final one is a spatio-temporal extension in the practical settings that allow spatial dependency of parameters to express nonstationary spatial behaviors. Let us begin this paper from the first one, the construction of block-orthogonal basis under irregular sampling.

2. Empirical Haar Wavelets

Let $s_p, p = 1, \ldots, n, \ldots$ be possibly countably infinite sampling points for spatial data, which we suppose are scattered irregularly over $[0,1]^2$. For bounded and real valued functions f and g defined over $[0,1]^2$, let us define the inner product by

$$< f, g >:= \lim_{n\to\infty} n^{-1} \sum_{p=1}^{n} f(s_p) g(s_p). \tag{1}$$

Under the inner product, Fourier basis is no more an orthogonal one except for the cases when sampling points are regularly spaced to provide mesh data.

We construct an orthogonal basis under the inner product. Let us start from the introduction of the one and two dimensional (1D and 2D) Haar wavelets that constitute an orthogonal basis over $[0,1]$ and $[0,1]^2$, respectively, under the Lebesgue measure. The mother wavelet over $[0,1]$ denoted as $\psi(x)$, which is shown in the upper of Figure 1, produces the orthogonal basis for $L^2([0,1])$ given by

$$\phi_{k,j}(x) := \psi(2^k x - j),$$

for $k = 0, 1, \ldots, j = 0, \ldots, 2^k - 1$, under the Lebesgue measure (see Theorem 1.1 of Ogden (1997)). The index k is called as a resolution parameter, since it controls the scale of the wavelet. The 1D Haar wavelets satisfy

$$\int_{[0,1]} \phi_{k,j}(x) \phi_{k',j'}(x) dx = c_k \delta_{kk'} \delta_{jj'}$$

for the positive constant $c_k = 4^{-k}$.

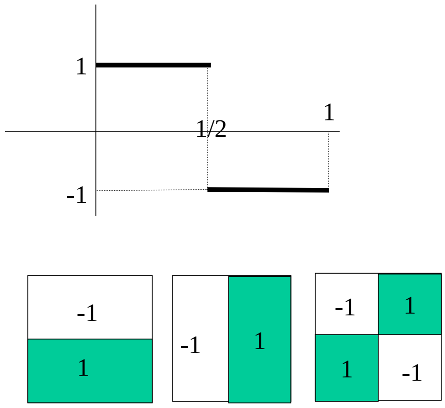


FIGURE 1. The mother wavelets for 1D and 2D Haar basis.

Similarly the mother wavelets on $[0,1]^2$, which are given by the three functions $\psi_1(x), \psi_2(x)$ and $\psi_3(x)$ defined over $[0,1]^2$ in lower part of Figure 1, produce the orthogonal basis on $L^2([0,1]^2)$ given by

$$\phi_{k,ij}(x_1,x_2) :=$$
$$\left(\psi_1(2^k x_1 - i, 2^k x_2 - j), \psi_2(2^k x_1 - i, 2^k x_2 - j), \psi_3(2^k x_1 - i, 2^k x_2 - j)\right), \tag{2}$$

for $k = 0, 1, \ldots, i, j = 0, \ldots, 2^k - 1$, under the Lebesgue measure (see Ogden(1997, page 167)). The index k also is called as a resolution parameter, since it controls the scale of the 2D wavelet. The 2D Haar wavelets satisfy

$$\int_{[0,1]^2} \phi'_{k,ij}(x_1,x_2)\phi_{k',i'j'}(x_1,x_2)dx_1dx_2 = C_k\delta_{kk'}\delta_{ii'}\delta_{jj'}$$

for the diagonal matrix $C_k = 4^{-k}I_3$.

Let us orthogonalize the 2D Haar wavelets $\phi_{k,ij}$ under the empirical measure in (1) for possibly countably infinite sampling points $s_p, p = 1, \ldots, n, \ldots$ scattered over $[0,1]^2$. Let $r(D)$ be the ratio of the number of the sampling points included in a subregion $D \subset [0,1]^2$ to that for $[0,1]^2$. We will orthogonalize $\phi_{k,ij}$ in the following three disjoint cases specified by the sampling points over its support $D_{k,ij}$. Let us denote the four disjoint equal-area sub-regions in $D_{k,ij}$ generated from the division by the lines $x_1 = (i + 0.5)/2^k$ and $x_2 = (j + 0.5)/2^k$ as $E^l_{k,ij}, l = 1, 2, 3, 4$.

Definition 1. *Let $D_{k,ij} \subset [0,1]^2$ be the region over which the Haar wavelet $\phi_{k,ij}$ is defined. We categorize the region $D_{k,ij}$ by the following three cases.*

Case 1 *All the four subregions $E^l_{k,ij}, l = 1, 2, 3, 4$ have positive ratios $r(E^l_{k,ij}) > 0$.*
Case 2 *Three of the four subregions have positive ratios.*
Case 3 *Two of the four subregions have positive ratios.*

Let us denote the two subregions of $D_{k,ij}$ over which the mth component of the Haar wavelet $\phi_{k,ij} = (\phi^1_{k,ij}, \phi^2_{k,ij}, \phi^3_{k,ij})$ takes values 1 and -1 as F_m and G_m, respectively, for $m = 1, 2, 3$. Modify the function values 1 and -1 of the original wavelet $\phi^m_{k,ij}$ to take $\sqrt{r(G_m)/r(F_m)}$ and $-\sqrt{r(F_m)/r(G_m)}$, respectively, and denote the modified function as $\tilde{\phi}^m_{k,ij}$ for $m = 1, 2, 3$. Let us define empirical Haar wavelets $\tilde{\phi}_{k,ij}$.

Definition 2. *The empirical Haar wavelet $\tilde{\phi}_{k,ij}$ is defined by depending on the category in Definition 1 as:*

(i) *In Case 1, $\tilde{\phi}_{k,ij}$ is defined by the three component function $(\tilde{\phi}^1_{k,ij}, \tilde{\phi}^2_{k,ij}, \tilde{\phi}^3_{k,ij})$.*
(ii) *In Case 2, $\tilde{\phi}_{k,ij}$ is defined by the two component function $(\tilde{\phi}^1_{k,ij}, \tilde{\phi}^2_{k,ij})$, where any two of $(\tilde{\phi}^1_{k,ij}, \tilde{\phi}^2_{k,ij}, \tilde{\phi}^3_{k,ij})$ can be chosen as the two components.*
(iii) *In Case 3, $\tilde{\phi}_{k,ij}$ is defined by the one component function $\tilde{\phi}^1_{k,ij}$, which is the one that satisfy that $r(F_m) > 0$ and $r(G_m) > 0$ from among the three components.*
(iv) *Otherwise $\tilde{\phi}_{k,ij}$ is not defined.*

Then a block orthogonality of $\tilde{\phi}_{k,ij}$ in Definition 2 is guaranteed under the empirical measure in (1).

Proposition 1. *For possibly countable infinite sampling points of $s_1, \ldots, s_n, \ldots,$ scattered irregularly over $[0, 1]^2$, the empirical Haar wavelets $\tilde{\phi}_{k,ij}$ in Definition 2 constitute a block orthogonal basis that satisfies*

$$\lim_{n\to\infty} n^{-1} \sum_{p=1}^{n} \tilde{\phi}'_{k,ij}(s_p) \tilde{\phi}_{k',i'j'}(s_p) = C_{k,ij} \delta_{kk'} \delta_{ii'} \delta_{jj'}$$

for the positive definite $m_{k,ij}$ by $m_{k,ij}$ matrix $C_{k,ij}$, where $m_{k,ij}$ is the number of components for $\tilde{\phi}_{k,ij}$ specified in Definition 2.

The proof is given in Section 7.

3. Spatial Data Modeling

3.1. **Wavelet model for spatial data.** We propose a model for spatial data by employing the empirical Haar wavelets $\tilde{\phi}_{k,ij}$ in Definition 2. Let us consider a modeling of a spatial process $Z(s)$ when sampling is conducted possibly on countably infinite points $s_p, p = 1, \ldots, n, \ldots$ over $[0, 1]^2$.

Let P_k be the set of (i, j)s that satisfy Cases 1, 2 or 3 in Definition 1 for a resolution k. We propose the following model for the spatial data by

$$Z(s_p) = \mu(s_p) + f(s_p) + \varepsilon_p, \tag{3}$$

where μ is a deterministic mean function, ε_p is a sequence of independent and identically distributed normal variables with mean 0 and variance σ^2 and f is the random process over $[0, 1]^2$ which is defined by

$$f(s) = \sum_{k=0}^{\infty} \sum_{(i,j)\in P_k} \tilde{\phi}_{k,ij}(s) \beta_{k,ij}, \tag{4}$$

where $\beta_{k,ij}$ is a sequence of independent normal vectors with mean 0 and variance

$$C(k + 1)^{-d} I_{m_{k,ij}}$$

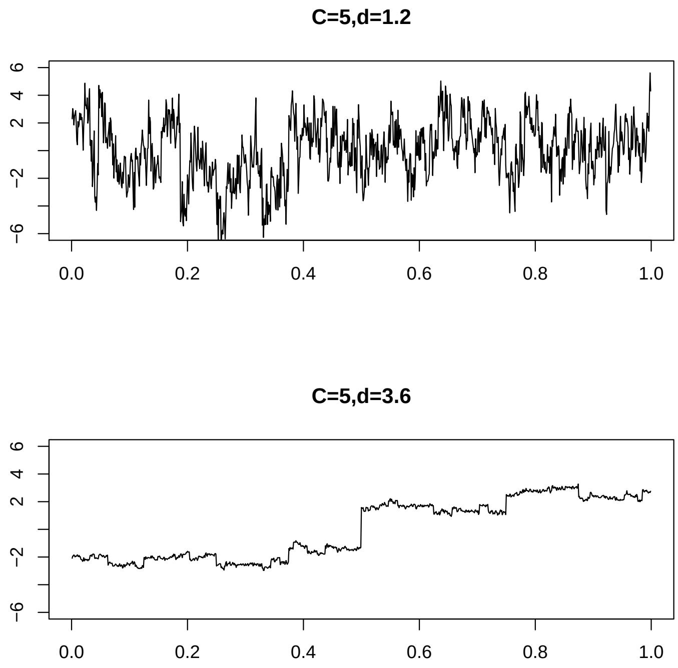


FIGURE 2. Sample paths of the 1D wavelet model in (3) for the cases of $(C, d) = (5, 1.2)$ and $(5, 3.6)$.

for $d > 1$, where $m_{k,ij}$ is the number of components for $\tilde{\phi}_{k,ij}$.

The wavelet model has nonstationary covariance function. By simple calculation, we have

$$\mathrm{Cov}\,(Z(s_p), Z(s_q)) = \sum_{k=0}^{\infty} \sum_{(i,j)\in P_k} \tilde{\phi}_{k,ij}(s_p)\tilde{\phi}'_{k,ij}(s_q)C(k+1)^{-d} + \sigma^2\delta_{pq},$$

which shows nonstationarity by the nature of the Haar wavelets.

In Figure 3.1, we show sample paths of the 1D wavelet model in (3) for $C = 5, d = 1.2$ and $C = 5, d = 3.6$ when $\mu = 0$ and $\sigma^2 = 0$. It is confirmed that larger d gives a smoother sample path because of more rapid decay of coefficients caused by the variance $C(k+1)^{-d}$. We call the parameter d as smoothness parameter since it controls the smoothness of the sample path by the wavelet model in (3).

3.2. **Whittle likelihood estimation.** We consider estimation of the parameters of C, d and σ^2 when the mean function $\mu(s)$ is absent in the wavelet model. Suppose we have observed $Z(s)$ that follows the model (3) on finite sampling points $s_1, \ldots, s_n$.

Let $\tilde{\phi}_{k,ij}$ be the n by $m_{k,ij}$ matrix given by $(\tilde{\phi}_{k,ij}(s_1), \ldots, \tilde{\phi}_{k,ij}(s_n))'$ and Z be the n by 1 vector given by $(Z(s_1), \ldots, Z(s_n))'$. The wavelet transform is defined by the following $m_{k,ij}$ by 1 vector,

$$w_{k,ij} = \left(\tilde{\phi}'_{k,ij}\tilde{\phi}_{k,ij}\right)^{-1}\tilde{\phi}'_{k,ij}Z, \tag{5}$$

which is equal to, by substituting the right hand side of (3) for Z and noting that the mean function is assumed to be absent,

$$\beta_{k,ij} + \left(\tilde{\phi}'_{k,ij}\tilde{\phi}_{k,ij}\right)^{-1}\tilde{\phi}'_{k,ij}\varepsilon, \tag{6}$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$. The block orthogonality of the empirical Haar wavelets makes the wavelet transform mutually orthogonal with mean 0 and variance

$$C(k+1)^{-d}I_{m_{k,ij}} + \sigma^2(\tilde{\phi}'_{k,ij}\tilde{\phi}_{k,ij})^{-1}, \tag{7}$$

which makes it easy and efficient to evaluate Gaussian likelihood for the wavelet transform, because the likelihood requires at most 3 dimensional operation. However the likelihood function is not sensitive to distinguish the first term from the second in (7) with respect to the parameters. The likelihood is not expected to work well for estimation. To remedy the inefficiency we slightly modify the form of the variance in (7). Choose two positive integers $p < q$ and redefine the variances for $k = 0, 1, \ldots, p$ as they are, and that for $k = q$ as the one obtained by dropping the first term in order to distinguish between the two terms. Namely, putting

$$\begin{aligned} W_{k,ij} &= w'_{k,ij}w_{k,ij}, \\ V_{k,ij}(d) &= (k+1)^{-d}I_{m_{k,ij}}, \\ B_{k,ij} &= \left(\tilde{\phi}'_{k,ij}\tilde{\phi}_{k,ij}\right)^{-1}, \end{aligned}$$

we propose the Gaussian likelihood function

$$\begin{aligned} &L(C, d, \sigma^2) = \\ &\sum_{k=0}^{p}\sum_{(i,j)\in P_k} \log\left|CV_{k,ij}(d) + \sigma^2 B_{k,ij}\right| + tr\left\{(CV_{k,ij}(d) + \sigma^2 B_{k,ij})^{-1}W_{k,ij}\right\} \\ &+ \sum_{(i,j)\in P_q} \log\left|\sigma^2 B_{q,ij}\right| + tr\left\{\left(\sigma^2 B_{q,ij}\right)^{-1}W_{q,ij}\right\}. \end{aligned} \tag{8}$$

The parameters C, d and σ^2 are estimated by the ones that minimize the likelihood in (8), which we call Whittle estimators in honor of the name for likelihood of Fourier transform in stationary time series.

3.3. **Consistency under fixed domain asymptotics.** We show the consistency of Whittle estimators for the wavelet model in (3) in the asymptotic situation as the sampling points $s_1, \ldots, s_n$ tends to infinity over the fixed region $[0,1]^2$. This is the kind of asymptotics called fixed domain asymptotics (Stein, 1999, page 62), since sampling points are denser and denser over the fixed region $[0,1]^2$ as n tends to be large.

To let the proof for the consistency easier to see, we redefine the parameters in (7) by

$$C\left\{(k+1)^{-d}I_{m_{k,ij}} + \tau(\tilde{\phi}'_{k,ij}\tilde{\phi}_{k,ij})^{-1}\right\}.$$

Concentrating C out of the Whittle likelihood yields the profile likelihood

$$l(d,\tau) =$$

$$\log A_n^{-1}\left[\sum_{k=0}^{p}\sum_{(i,j)\in P_k} tr\left\{(V_{k,ij}(d)+\tau B_{k,ij})^{-1}W_{k,ij}\right\}+\sum_{(i,j)\in P_q} tr\left\{(\tau B_{q,ij})^{-1}W_{q,ij}\right\}\right]$$

(9)

$$+A_n^{-1}\left\{\sum_{k=0}^{p}\sum_{(i,j)\in P_k}\log|V_{k,ij}(d)+\tau B_{k,ij}|+\sum_{(i,j)\in P_q}\log|\tau B_{q,ij}|\right\},$$

where

$$A_n=\sum_{k=0}^{p}\sum_{(i,j)\in P_k} m_{k,ij}+\sum_{(i,j)\in P_q} m_{q,ij}.$$

The followings are assumed to prove the consistency.

Assumption 1. *For the wavelet model in (3), we assume:*

(1) *The wavelet model has no mean function.*
(2) *The sequences of random variables ε_p and $\beta_{k,ij}$ in the wavelet model are Gaussian.*
(3) *The parameter spaces for d and τ are $D=(1,\delta_0]$ and $\Theta=[\delta_1,\delta_2]$, which include d_0 and τ_0 as true values inside the regions, respectively, for $\delta_0>1$ and $\delta_1>\tau_0/2$.*
(4) *There exist sequences of $p=p_n$ and $q=q_n$ with $p_n<q_n$ that diverge as ntends to be large such that*

$$\sup_{(i,j)\in P_p} p^{\delta_0}\left|\left|\left(\tilde{\phi}'_{p,ij}\tilde{\phi}_{p,ij}\right)^{-1}\right|\right|_2\to 0,$$

$$\left(\sum_{(i,j)\in P_q} m_{q,ij}\right)^{-1}\left(\sum_{k=0}^{p}(k+1)^{3/2-\delta_0}\sqrt{\sum_{(i,j)\in P_k} m_{k,ij}}\right)\to 0,$$

as n tends to be large.

Consistency for the Whittle estimators

$$(\hat{d},\hat{\tau})=\arg\min_{d\in D,\tau\in\Theta} l(d,\tau)$$

and

$$\hat{C}=A_n^{-1}\left[\sum_{k=0}^{p}\sum_{(i,j)\in P_k} tr\left\{(V_{k,ij}(\hat{d})+\hat{\tau}B_{k,ij})^{-1}W_{k,ij}\right\}+\sum_{(i,j)\in P_q} tr\left\{(\hat{\tau}B_{q,ij})^{-1}W_{q,ij}\right\}\right]$$

are shown in the following Theorem 1.

Theorem 1. *Under Assumption 1, Whittle estimators $\hat{d}$, $\hat{\tau}$ and $\hat{C}$ converge in probability to d_0, τ_0 and C_0, respectively, as n tends to be infinity, and hence $\hat{\sigma}^2=\hat{C}\hat{\tau}$ is also consistent.*

3.4. **Kriging.** Suppose we have identified the wavelet model in (3) for samples over $s_1, \ldots, s_n$. Let us show the way to estimate the value of an unsampled point $u \in [0,1]^2$, which is called traditionally as kriging. Assume that the mean function is 0 without loss of generality for simplicity. Our model provides an efficient way for kriging that requires at most 3 dimensional matrix operation.

Bayes theorem applies to kriging in the wavelet model (3), when we regard the distribution for $\beta_{k,ij}$ as a prior. The wavelet transform $w_{k,ij}$ is from (6) normally distributed independently with mean $\beta_{k,ij}$ and variance $\sigma^2(\tilde{\phi}'_{k,ij}\tilde{\phi}_{k,ij})^{-1}$, where the parameter $\beta_{k,ij}$ has the normal prior with mean 0 and variance $C(k+1)^{-d}I_{m_{k,ij}}$. By Bayes Theorem, the posterior variance and mean are given by

$$\begin{aligned} v_{k,ij}^{-1} &= C^{-1}(k+1)^d I_{m_{k,ij}} + \left(\tilde{\phi}'_{k,ij}\tilde{\phi}_{k,ij}\right)/\sigma^2, \\ b_{k,ij} &= v_{k,ij}\left(\tilde{\phi}'_{k,ij}\tilde{\phi}_{k,ij}\right) w_{k,ij}/\sigma^2, \end{aligned}$$

respectively. It follows that the posterior distribution for $Z(u)$ is normal with mean and variance given by

$$\begin{aligned} &\sum_{k=0}^{\infty}\sum_{(i,j)\in P_k} \tilde{\phi}_{k,ij}(u) b_{k,ij}, \\ &\sum_{k=0}^{\infty}\sum_{(i,j)\in P_k} \tilde{\phi}_{k,ij}(u) v_{k,ij}\tilde{\phi}'_{k,ij} + \sigma^2, \end{aligned} \tag{10}$$

respectively. The mean square error as well as kriging are obtained efficiently with at most 3 dimensional matrix operations.

4. Spatio-Temporal Extension

4.1. **Wavelet model for spatio-temporal data.** We shall extend the wavelet model in (3) for spatial data to that for spatio-temporal data. Suppose that, for temporal points $t = 1, \ldots, T$, $Z(t,s)$ is onserved on $s_{t,m}, m = 1, \ldots, n_t$ that are scattered irregularly over $[0,1]^2$. We allow $s_{t,p}$ to depend on t to account for the feature that sampling points may not be the same at each temporal point. Let S_t be the set of the sampling points for $t = 1, \ldots, T$ and define S_0 by $\cap_{t=1}^T S_t$. Let P_k be the set of (i,j) that is included in any of Cases 1-3 in Definition 1 for the set S_0 and $k = 0, 1, \ldots$. For each t, we construct the empirical Haar wavelets for $(i,j) \in P_k$, k=0,1,2,..., by the ones that constitute orthogonal basis for $s \in S_t$. The empirical Haar wavelets depends on t, since S_t depends on t. To show the dependency explicitly, we denote the empirical Haar wavelet for each t as $\tilde{\phi}_{k,ij}(t,s)$.

We fit the spatial model in (3) to $Z(t, s_{t,m})$ for each t and conduct temporal extension by regarding the regression coefficients $\beta_{k,ij}$ as the state vector that follows autoregressive models. We propose the state space model for spatio-temporal data by

$$Z(t, s_{t,m}) = \mu(t, s_{t,m}) + f(t, s_{t,m}) + \varepsilon_m(t),$$

for $t = 1, \dots, T$ and $m = 1, \dots, n_t$, where μ is a mean function, and

$$f(t,s) = \sum_{k=0}^{\infty} \sum_{(i,j)\in P_k} \tilde{\phi}_{k,ij}(t,s)\beta_{k,ij}(t), \tag{11}$$
$$\beta_{k,ij}(t+1) = \rho_{k,ij}\beta_{k,ij}(t) + u_{k,ij}(t),$$

for $t = 1, \dots, T$, where $u_{k,ij}(t)$ is a normal random vector with mean 0 and variance matrix

$$A_{k,ij} I_{m_{k,ij}},$$

for $|\rho_{k,ij}| < 1$.

The model in (11) can express non-separable space time covariance functions. By simple calculations, we have

$$\begin{aligned} &\operatorname{cov}\left(Z(t,u), Z(t-h,v)\right) \\ &= \sum_{(k,i,j)\in P_0} \tilde{\phi}_{k,ij}(t,u)\tilde{\phi}'_{k,ij}(t-h,v)A_{k,ij}\rho_{k,ij}^h + \sigma^2\delta_{uv}\delta_{t,t-h}, \end{aligned} \tag{12}$$

from which the spatial and temporal covariances are not separated unless $\rho_{k,ij}$ is a constant.

4.2. **Parametric modeling of space time covariance functions.** Allowing C, d in the spatail model in (3) to depend on k, i, j, we identify $A_{k,ij}$ by

$$A_{k,ij} = C_{k,ij}(k+1)^{-d_{k,ij}},\ d > 1. \tag{13}$$

Points are in how $C_{k,ij}, d_{k,ij}$ and $\rho_{k,ij}$ are described by parametric functions.

The dependencies of C, d and ρ on k, i, j are interpreted as the local dependencies of the scale, smoothness and temporal correlation in the space time covariance structures in (12). It is of crucial importance how the dependencies are identified in parametric forms to express locally dependent spatial behaviors.

We shall identify $C_{k,ij}, d_{k,ij}$ and $\rho_{k,ij}$ with functions f_1, f_2 and f_3 defined over $[0,1]^2$ through the integral averages

$$|D_{k,ij}|^{-1} \int_{D_{k,ij}} f_m(x)dx, \tag{14}$$

for $m = 1, 2$ and 3, respectively, where $|D_{k,ij}|$ is the area of the domain $D_{k,ij}$. To describe the spatial functions in a parametric form, we employ a Fourier expansion

$$f_m(x) = a_0 + \sum_{p\in Q} \left\{a_{m,p}\cos(\omega_p' x) + b_{m,p}\sin(\omega_p' x)\right\}, \tag{15}$$

where ω_p is a frequency which we suggest to design by $(2\pi i_p, 2\pi j_p)'$ for the set Q of mesh points (i_p, j_p) except for the origin over the upper half plane of $\mathbf{R}^2$. The underlying functions f_1, f_2 and f_3 are interpreted as the locally dependent scale, smoothness and temporal correlation, respectively.

4.3. **Whittle likelihood estimation.** This section considers estimation for parameters that describe $A_{k,ij}$ and $\rho_{k,ij}$ by the parametric models in (15). To show explicitly the dependency on parameters, we express the functions as $A_{k,ij}(\theta)$ and $\rho_{k,ij}(\theta)$ for parameters $\theta \in \Theta$. We assume $\mu = 0$ for simplicity.

Exactly as in (5), we define the wavelet transform by

$$w_{k,ij}(t) = \left(\tilde{\phi}'_{k,ij}(t)\tilde{\phi}_{k,ij}(t)\right)^{-1} \tilde{\phi}'_{k,ij}(t)Z(t),$$

for $t = 1, \dots, T$, where $\tilde{\phi}'_{k,ij}(t) = (\tilde{\phi}'_{k,ij}(t, s_{t,1}), \dots, \tilde{\phi}'_{k,ij}(t, s_{t,n_t}))'$ and $Z(t) = (Z(t, s_{t,1}), \dots, Z(t, s_{t,n_t}))'$.

By the block orthogonality of $\tilde{\Phi}_{k,ij}(t)$, the state space model for $Z(t, s_{t,p})$ in (11) reduces to, for $(i, j) \in P_k, k = 0, 1, \dots, p$,

$$\begin{aligned} w_{k,ij}(t) &= \beta_{k,ij}(t) + f_{k,ij}(t), \\ \beta_{k,ij}(t+1) &= \rho_{k,ij}(\theta)\beta_{k,ij}(t) + u_{k,ij}(t), \end{aligned} \tag{16}$$

where $f_{k,ij}(t)$ is the independent random vector with mean 0 and variance matrix $\sigma^2(\tilde{\phi}'_{k,ij}(t)\tilde{\phi}_{k,ij}(t))^{-1}$, which is independent for each k, i, j. It follows that the likelihood function for the wavelet transform, which we call Whittle likelihood again, is efficiently evaluated by following Kalman recursion.

Let us introduce an explicit form of Kalman recursion to evaluate the likelihood. Put

$$\begin{aligned} \mu_{k,ij}(t) &= E(\beta_{k,ij}(t)|w_{k,ij}(t), \dots, w_{k,ij}(1)), \\ Q_{k,ij}(t) &= \mathrm{var}(\beta_{k,ij}(t)|w_{k,ij}(t), \dots, w_{k,ij}(1)), \\ \eta_{k,ij}(t+1) &= E(\beta_{k,ij}(t+1)|w_{k,ij}(t), \dots, w_{k,ij}(1)), \\ P_{k,ij}(t+1) &= \mathrm{var}(\beta_{k,ij}(t+1)|w_{k,ij}(t), \dots, w_{k,ij}(1)), \end{aligned}$$

which are evaluated by the Kalman recursion when we initialize by $\eta_{k,ij}(1) = O_{m_{k,ij}}$ and $P^{-1}_{k,ij}(1) = O_{m_{k,ij}}$, the diffuse prior, namely by

$$\begin{aligned} Q^{-1}_{k,ij}(t) &= P^{-1}_{k,ij}(t) + \sigma^{-2}\left(\tilde{\phi}'_{k,ij}(t)\tilde{\phi}_{k,ij}(t)\right), \\ \mu_{k,ij}(t) &= Q_{k,ij}(t)\left\{P^{-1}_{k,ij}(t)\eta_{k,ij}(t) + \sigma^{-2}\left(\tilde{\phi}'_{k,ij}(t)\tilde{\phi}_{k,ij}(t)\right) w_{k,ij}(t)\right\}, \\ \eta_{k,ij}(t+1) &= \rho_{k,ij}\mu_{k,ij}(t), \\ P_{k,ij}(t+1) &= \rho^2_{k,ij}Q_{k,ij}(t) + A_{k,ij}I_{m_{k,ij}}, \\ F_{k,ij}(t+1) &= P_{k,ij}(t+1) + \sigma^2\left(\tilde{\phi}'_{k,ij}(t+1)\tilde{\phi}_{k,ij}(t+1)\right)^{-1}, \end{aligned} \tag{17}$$

for $t = 1, \dots, T$ and $(i, j) \in P_k, k = 0, 1, 2, \dots, p$.

Define the $m_{k,ij} \times m_{k,ij}$ periodogram matrix by

$$J_{k,ij}(t) := (w_{k,ij}(t) - \eta_{k,ij}(t,\theta))(w_{k,ij}(t) - \eta_{k,ij}(t,\theta))'.$$

Then the likelihood function for the wavelet transform is evaluated as

$$\log L_{ww}(\theta) := \sum_{t=2}^{T}\sum_{k=0}^{p}\sum_{(i,j)\in P_k}\left\{tr\left(J_{k,ij}(t)F^{-1}_{k,ij}(t,\theta)\right) + \log|F_{k,ij}(t,\theta)|\right\}, \tag{18}$$

which we call the wavelet version of Whittle likelihood for spatio-temporal data. The parameter θ that describes $A_{k,ij}$ and $\rho_{k,ij}$ by (15) is estimated by minimizing Whittle likelihood.

4.4. **Asymptotic distribution.** Asymptotic distribution for Whittle estimators minimizing (18) is derived when temporal size T tends to be large with finitely fixed spatial sampling points. Unlike Whittle estimators for the spatial model in (3) under fixed domain asymptotics, traditional theories for maximum likelihood estimation can apply to have the asymptotic distribution. Assume that spatio-temporal data follows (11) with $\sup_{k=0,1,2,\dots,}\sup_{(i,j)\in P_k}|\rho_{k,ij}|<1$. Let θ_0 be the true value for the parameters θ. Then temporal stationarity of the state vector processes $\beta_{k,ij}(t)$ in (16) is guaranteed for $k=0,\dots,p,(i,j)\in P_k$, and the distribution of $\hat{\theta}$ minimizing (18) is approximately, as T tends to be large,

$$N\left(\theta_0,\Omega(\theta_0)\right),$$

where

$$\Omega(\theta)=\left(\frac{1}{2}\frac{\partial^2\log L_{ww}(\theta)}{\partial\theta\partial\theta'}\right)^{-1}.$$

4.5. **Forecasting.** Following the spatio-temporal data analysis literatures, we denote estimation for future values at $t>T$ as forecasting. Forecasting is conducted efficiently as a result of the Kalman recursion in (17). For $h=1,2,\dots$, define

$$\eta^h_{k,ij}(T)=\rho^h_{k,ij}\mu_{k,ij}(T),$$
$$P^h_{k,ij}(T)=\rho^{2h}_{k,ij}Q_{k,ij}(T)+\sum_{a=0}^{h-1}\rho^{2a}A_{k,ij}I_{m_{k,ij}}$$

and the forecasting for a spatial point u at time $T+h$ is

$$\sum_{k=0}^{\infty}\sum_{(i,j)\in P_k}\tilde{\phi}_{k,ij}(T,u)\eta^h_{k,ij}(T+1) \tag{19}$$

for which the mean square error is evaluated as

$$\sum_{k=0}^{\infty}\sum_{(i,j)\in P_k}\tilde{\phi}_{k,ij}(T,u)P^h_{k,ij}(T)\tilde{\phi}'_{k,ij}(T,u)+\sigma^2.$$

5. Empirical examples

This section examines the empirical performances of the wavelet modeling by applying to real data. Public land price data in Tokyo, which is composed of public land prices for sampling points irregularly scattered over Tokyo, Kanagawa, Chiba and Saitama prefectures, are employed for the demonstration.

Let us summarize briefly public land price data in Japan. Government of Japan has been publicizing land prices in the form of price per m^2 as of January 1st over hundred thousands of sampling points scattered irregularly all over Japan. They evaluate land prices by combinations of real record of transactions, income approach and cost accounting method, which are therefore not necessarily equal to real market prices. The purpose of the publications in March every year since 1970 in the government web page is to help evaluate public works, inheritance tax and so on.

This paper focuses on yearly series of land prices in Tokyo areas from 2003 till 2014 to demonstrate efficiency of our wavelet approach. The yearly series of land prices has the following features. First one is large number of sampling points, which is about 6000-8000 at each temporal point and amounts to 100,000 overall.

year	03	04	05	06	07	08	09	10	11	12	13	14
no.	8554	8554	8378	8348	8025	7829	7509	7360	6881	6881	6878	5977

TABLE 1. Number of sampling points for public land prices in Tokyo area from 2003 to 2014.

Sampling points in 2003 are shown in Figure 5, fro which irregularity of sampling points is confirmed. The second one is that sampling points at each temporal point are not necessarily the same year by year. Table 1 shows the actual number of sampling points in each year. The final one is clear exhibition of spatial nonstationarity between behaviors in central Tokyo and in suburbs of Tokyo. The three features make it impossible to apply existing approaches such as vector ARMA models in a trivial way and provide us with a good chance to demonstrate our proposed approach by the empirical Haar wavelet expansions.

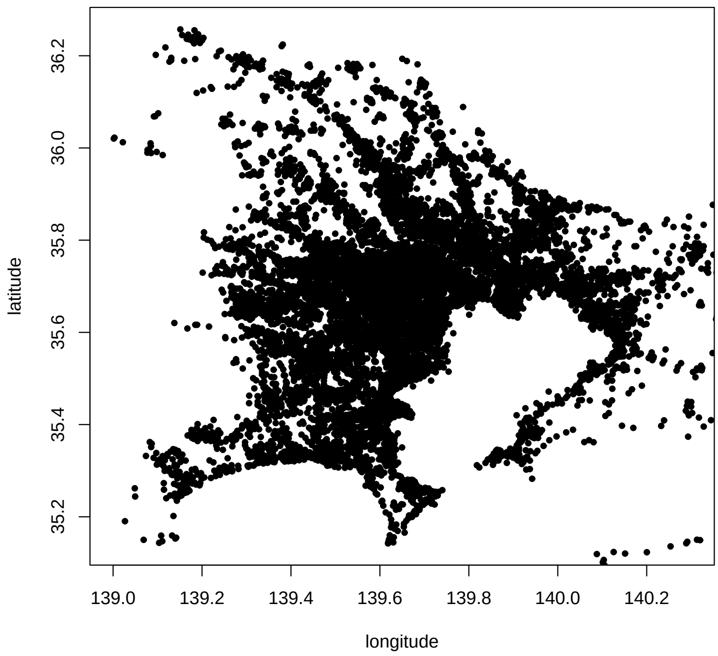


FIGURE 3. 8554 sampling points for public land price data in Tokyo in 2003.

Let $P_t(s)$ be the land price at spatial point s at temporal point t. Spatio-temporal data of log return that is the first differences of log prices, namely

$$r_t(s_i) = \log P_t(s_i) - \log P_{t-1}(s_i), \ t = 2004, \dots, 2014, \tag{20}$$

is our object for analysis, where s_i is a ith sampling point that has observations at two consecutive years $t-1$ and t.

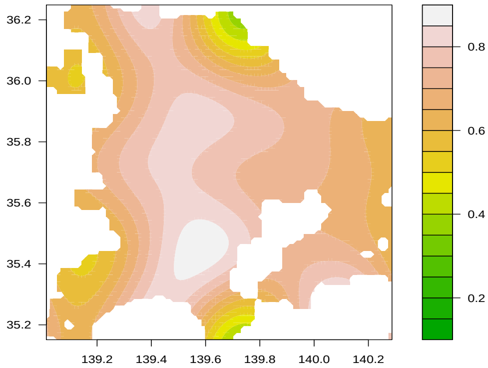
year	$\hat{d}$	$\hat{\sigma}^2$	S/N	wavelet	wavelet thresholding	
				MSE	thrshld=0	$\sqrt{\log n}\sigma$
2004	3.01	3.06	8.17	5.47	7.56	6.40
2005	2.43	2.59	6.07	3.34	6.26	4.13
2006	1.36	2.17	3.94	2.15	2.61	3.55
2007	1.23	3.60	5.50	5.61	9.39	9.22
2008	1.15	4.37	5.62	2.69	5.08	5.50
2009	1.65	1.45	2.45	1.31	1.93	3.34
2010	1.46	1.71	2.76	1.31	1.81	3.19
2011	1.43	0.79	1.65	0.67	1.47	1.48
2012	1.78	0.61	2.60	0.59	1.05	0.77
2013	1.46	0.47	2.60	0.47	0.95	0.78
2014	1.72	0.66	2.75	0.66	1.61	2.74

TABLE 2. Whittle estimatots for d and σ^2 in (3) with the SN ratios and MSEs of the wavelet kriging in comparison with those of wavelet thresholding, where MSEs were evaluated for randomly chosen 100 points among the samples in each year.

First we fit the wavelet model for spatial data in (3) to the log return in $t = 2004, \ldots, 2014$. We conduct Whittle estimation for the parameters C, d, σ^2, minimizing Whittle likelihood with $p = 6, q = 7$ in (8), where the samples used for evaluating kriging mean square errors (MSEs) mentioned just below are removed. With the estimated parameters, we conduct kriging in each year by the formula in (10) and evaluate MSEs of the kriging for randomly chosen 100 points from among samples. Estimated results as well as MSEs of the krigings are shown in Table 2, where MSEs of wavelet thresholding (Donoho and Johnstone, 1994) are also shown as benchmarks. Wavelet thresholding is a kriging just by substituting $w_{k,ij}$ for $\beta_{k,ij}$ in (3) when $|w_{k,ij}|$ is larger than a threshold value and substituting 0 otherwise. Here the threshold is specified with 0 or $\sqrt{\log n}\sigma$ which is motivated by certain properties of the normal distribution discussed in detail by Donoho and Johnstone (1994). Our method outperforms the benchmarks in terms of MSEs every year. The SN ratio, which is $var(f(s))/var(\varepsilon)$ in (3), declined suddenly in 2009 and has been held low after Lehman shock. Interpreting $var(f(s))$ as general spatial tendency reflecting market activity, we find that land market activity in Tokyo became shrink just after the shock and has not yet recovered yet even in 2014.

Next we fit the wavelet model for spatio-temporal datal in (11) to the yearly series of log return in (20) separately before and after Lehman shock occurred in September of 2008, namely, separately to the series from 2004 to 2008 and to that from 2009 to 2013, where samples in 2014 were excluded for use of validating forecasting performances later. We will focus on the cases when $C_{k,ij}, d_{k,ij}$ in (13) and $\rho_{k,ij}$ that describes the temporal correlations are specified by (14), where we set $Q = \{(i,j)||i| \leq 2, 0 \leq j \leq 2\} \cap \{(i,j)|i > 0 \text{ or } j > 0\}$ in (15). The identified model has the 146 parameters for the data size of 75,166 as a result.

We conduct Whittle estimation for the parameters by minimizing the likelihood in (18) with $p = 7$. We show the figures of the underlying functions in (14) for $\rho_{k,ij}$ only as the ones that identify interesting features of land prices before and after Lehman shock. It is observed after Lehman shock that the temporal correlations declined to be nearly independent in central Tokyo, while they hold high in suburbs of Tokyo.



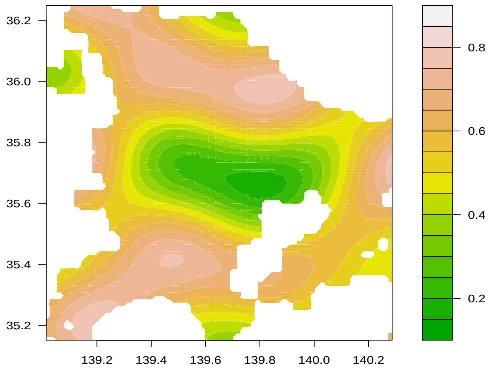


FIGURE 4. The identified temporal correlation structures before and afyer Lehman shock, the underlying functions in (14) for $\rho_{k,ij}$.

To see the goodness of fit for the identified spatio-temporal model, we construct forecasts by the formula in (19) for randomly chosen 500 points from samples in 2014 and for points chosen from among them over which temporal correlations have been identified as higher than .6 and lower than .3. For comparisons, we consider a benchmark which forecasts always 0. In Tabke 3, MSEs for both forecasts are shown. It is found that our wavelet method works better especially in highly correlated areas and hence it detects well temporal correlation structures and provides reasonable forecasts based on the identified correlations.

6. DISCUSSION

This paper proposes a model for irregularly spaced data by orthogonal expansion with Haar wavelets and conducts its spatio-temporal extension. The spatio-temporal model can express nonseparable and spatially nonstationary behaviors in

	no. samples	wavelet	null forecast
overall	500	2.79	3.65
correl.$>.6$	165	1.31	2.43
correl.$<.3$	133	6.33	6.93

Table 3. MSEs of the forecasts by the wavelet model and benchmark that predicts always 0 for randomly chosen 500 points from samples in 2014 and for the points from among them over which temporal correlations are identified with higher than .6 and with lower than .3. The forecastsl are constructed with the model identified by the samples till 2013.

space time covariances by allowing the prior variances of the wavelet coefficients dependent spatially. The crucial contribution is the discovery of a way to construct orthogonal basis under irregular sampling via Haar wavelets, which makes our approach unique in the literatures for spatial or spatio-temporal data analysis. The orthogonal expansion by the empirical Haar wavelets provides us with efficient ways for kriging or forecasting as well as parameter estimation that requires at most three dimensional matrix operations, which opens a way to analyze huge spatially nonstationary data set of several hundred thousand space time points.

We state some points that are to be studied in the futures. First one is a temporal nonstationary extension of the stationary temporal covariances that this paper assumed in the nonseparable space time covariances in (12). In order to allow nonstationary in the temporal dimension, there are two possible ways. One is to apply random walk model instead of the autoregressive model to the state vectors in (11). This approach is expected to be effective for nonstationary data in both space and time when sampling is conducted regularly in time but irregularly in space, which are popular in social science areas. The other one is the use of three dimensional Haar wavelets to construct an orthogonal basis under irregular sampling over three dimensional space. Then the model by the empirical three dimensional Haar wavelets describes nonstationary covariances jointly in both space and time. This approach is expected to be effective for spatio-temporal data collected irregularly in both space and time, which frequently appears in natural science areas.

Second one is possibility of jointly conducting estimation and kriging by full-Bayes approach. The models proposed in the paper are Bayesian regression models whose regression coefficients have prior distributions. Our approach is considered as an empirical Bayes method that estimates hyper parameters by marginal likelihood functions, which is Whittle likelihood in our case, and conducts kriging or forecasting separately from the estimation. The jointly conducting approach by full-Bayes method with MCMC sampling can improve the performances of our empirical Bayes approach, though it is much more time consuming.

Final one is a multivariate extension of the univariate wavelet model, which we expect is possible by a combination with the method of coregionalization (Banerjee et. al. (2004)). It would make it possible to analyze a case when all components of multivariate observations are not necessarily observed.

7. PROOF

7.1. **proof of Proposition 1.** It is trivial by simple algebra that, for $(k,i,j) \neq (k',i',j')$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{p=1}^{n} \tilde{\phi}_{k,ij}(s_p)\tilde{\phi}_{k',i'j'}(s_p) = 0.$$

We will show that $C_{k,ij}$ is positive definite in Case 1. Put $R(E^l_{k,ij}) = r_l$ for $l = 1,2,3,4$. Let us prove by contradiction. Assume that $C_{k,ij}$ is singular. Since $\tilde{\phi}_{k,ij}$ is linearly dependent, there exist a non-zero constants k_1, k_2 and k_3 that satisfy

$$\begin{aligned}
k_1\frac{1}{r_1+r_2} + k_2\frac{1}{r_1+r_3} + k_3\frac{1}{r_1+r_4} &= 0\\
k_1\frac{1}{r_1+r_2} - k_2\frac{1}{r_2+r_4} - k_3\frac{1}{r_2+r_3} &= 0\\
-k_1\frac{1}{r_3+r_4} + k_2\frac{1}{r_1+r_3} - k_3\frac{1}{r_2+r_3} &= 0\\
-k_1\frac{1}{r_3+r_4} - k_2\frac{1}{r_2+r_4} + k_3\frac{1}{r_1+r_4} &= 0.
\end{aligned}$$

It follows that the matrix given by

$$B = \begin{bmatrix} \frac{1}{r_1+r_2} & \frac{1}{r_1+r_3} & \frac{1}{r_1+r_4} \\ \frac{1}{r_1+r_2} & -\frac{1}{r_2+r_4} & -\frac{1}{r_2+r_3} \\ -\frac{1}{r_3+r_4} & \frac{1}{r_1+r_3} & -\frac{1}{r_2+r_3} \end{bmatrix}$$

has a rank that must be less than 3. By elementary row operations, the matrix B reduces to the upper triangular matrix given by

$$\begin{bmatrix} \frac{1}{r_1+r_2} & \frac{1}{r_1+r_3} & \frac{1}{r_1+r_4} \\ 0 & -\frac{r_1+r_2+r_3+r_4}{(r_2+r_4)(r_1+r_3)} & -\frac{r_1+r_2+r_3+r_4}{(r_1+r_4)(r_2+r_3)} \\ 0 & 0 & -\frac{2r_4(r_1+r_2+r_3+r_4)}{(r_1+r_4)(r_2+r_3)} \end{bmatrix}.$$

r_4 must be 0 to let the matrix B be singular, which contradicts the assumption that r_1, r_2, r_3, r_4 must be all positive.

In Cases 2 and 3, the positive definiteness is proved more easily by the same argument.

7.2. **proof of Theorem 1.** We will employ the method by Hurvich et. al (2005) who proved the consistency of the long memory parameter estimation in long memory time series by by local Whittle estimation.

7.2.1. *Consistency of $\hat{d}$.* Replacing $W_{k,ij}$ and $W_{q,ij}$ in (9) with $V_{k,ij}(d_0) + \tau_0 B_{k,ij}$ and $\tau_0 B_{q,ij}$, respectively, we define $l_0(d,\tau)$ by

$$\begin{aligned}
l_0(d,\tau) = &\log A_n^{-1}\left[\sum_{k=0}^{p}\sum_{(i,j)\in P_k}(k+1)^{d-d_0} tr\left\{\alpha_{k,ij}(d,\tau)\right\} + \tau^{-1}\tau_0 \sum_{(i,j)\in P_q} m_{q,ij}\right]\\
&+ A_n^{-1}\left\{\sum_{k=0}^{p}\sum_{(i,j)\in P_k}\log\left|(k+1)^{-d}H_{k,ij}(d,\tau)\right| + \sum_{(i,j)\in P_q}\log|\tau B_{q,ij}|\right\},
\end{aligned} \tag{21}$$

where

$$\begin{aligned} H_{k,ij}(d,\tau) &= I_{m_{k,ij}} + \tau(k+1)^d B_{k,ij}, \\ \alpha_{k,ij}(d,\tau) &= H_{k,ij}^{-1}(d,\tau) H_{k,ij}(d_0,\tau_0). \end{aligned}$$

Since (d_0, τ_0) minimizes l_0,

$$\begin{aligned} 0 &\leq l_0(\hat{d},\hat{\tau}) - l_0(d_0,\tau_0) \\ &= l_0(\hat{d},\hat{\tau}) - l(\hat{d},\hat{\tau}) + l(\hat{d},\hat{\tau}) - l(d_0,\tau_0) + l(d_0,\tau_0) - l_0(d_0,\tau_0) \\ &\leq 2 \sup_{(d,\tau)\in D\times\Theta} |l(d,\tau) - l_0(d,\tau)| \\ &= 2 \sup_{(d,\tau)\in D\times\Theta} \left| \log \left(1 + \sum_{k=0}^{p} \sum_{(i,j)\in P_k} \gamma_{k,ij}(d,\tau) + \sum_{(i,j)\in P_q} \gamma_{q,ij}(d,\tau) \right) \right|, \end{aligned}$$

where

$$\begin{aligned} \gamma_{k,ij}(d,\tau) &= \frac{(k+1)^{d-d_0} tr \left\{ \alpha_{k,ij}(d,\tau) \left(E_{k,ij} - I_{m_{k,ij}} \right) \right\}}{M_n(d,\tau)}, \\ \gamma_{q,ij}(d,\tau) &= \frac{\tau^{-1}\tau_0 tr \left(E_{q,ij} - I_{m_{q,ij}} \right)}{M_n(d,\tau)}, \\ M_n(d,\tau) &= \sum_{k=0}^{p} \sum_{(i,j)\in P_k} (k+1)^{d-d_0} tr\left(\alpha_{k,ij}(d,\tau)\right) + \tau^{-1}\tau_0 \sum_{(i,j)\in P_q} m_{q,ij}, \\ E_{k,ij} &= C_0^{-1} k^{d_0} H_{k,ij}^{-1}(d_0,\tau_0) W_{k,ij}, \\ E_{q,ij} &= C_0^{-1} \tau_0^{-1} B_{q,ij}^{-1} W_{q,ij}. \end{aligned}$$

Define

$$\begin{aligned} K_n(x,\tau^{-1}\tau_0) = & \log A_n^{-1} \left\{ \sum_{k=0}^{p} (k+1)^x \sum_{(i,j)\in P_k} m_{k,ij} + \tau^{-1}\tau_0 \sum_{(i,j)\in P_q} m_{q,ij} \right\} \\ & - A_n^{-1} \left\{ x \sum_{k=0}^{p} \log(k+1) \sum_{(i,j)\in P_k} m_{k,ij} + \log(\tau^{-1}\tau_0) \sum_{(i,j)\in P_q} m_{q,ij} \right\}. \end{aligned}$$

The function $K_n(x,y)$ is a non-negative function and twice differentiable with respect to x with a positive second derivative that is bounded away from 0 for all $(d,\tau) \in D \times \Theta$, and there exists x_0 such that $\partial K(x_0,y)/\partial x = 0$ for a fixed y. Thus there exist positive constants c_1 and c_2 such that for $(d,\tau) \in D \times \Theta$,

$$\begin{aligned} K_n(d-d_0,\tau^{-1}\tau_0) &= K_n(x_0,\tau^{-1}\tau_0) + \frac{\partial^2 K_n(\tilde{x},\tau^{-1}\tau_0)}{\partial x^2}(d-d_0-x_0)^2 \\ &\geq K_n(x_0,\tau^{-1}\tau_0) + c_1(d-d_0-x_0)^2 \geq c_2(d-d_0)^2, \end{aligned}$$

where in the last inequality we used $x_0 = 0$ when $\tau^{-1}\tau_0 = 1$ and that $K_n(x,y)$ is non-negative, continuous and attains 0 if and only if $x = 0, y = 1$. Hence, defining

$R_n(d,\tau) = l_0(d,\tau) - l_0(d_0,\tau_0) - K_n(d-d_0,\tau^{-1}\tau_0)$, we obtain

$$\begin{aligned}
0 &\le (\hat{d}-d_0)^2 1_D(\hat{d}) \le c_2^{-1} K_n(\hat{d}-d_0,\tau^{-1}\tau_0) \\
&\le c_2^{-1} l_0(\hat{d},\tau) - c_2^{-1} l_0(d_0,\tau_0) - c_2^{-1} R_n(\hat{d},\tau) \\
&\le 2c_2^{-1} \sup_{(d,\tau)\in D\times\Theta} \left| \log\left(1 + \sum_{k=0}^{p} \sum_{(i,j)\in P_k} \gamma_{k,ij}(d,\tau) + \sum_{(i,j)\in P_q} \gamma_{q,ij}(d,\tau)\right)\right| \\
&+ c_2^{-1} \sup_{(d,\tau)\in D\times\Theta} |R_n(d,\tau)|,
\end{aligned}$$

and

$$\begin{aligned}
R_n(d,\tau) =& \log\left\{1 + \frac{\sum_{k=0}^{p}(k+1)^{d-d_0}\sum_{(i,j)\in P_k} tr\left(\alpha_{k,ij}(d,\tau) - I_{m_{k,ij}}\right)}{\sum_{k=0}^{p}(k+1)^{d-d_0}\sum_{(i,j)\in P_k} m_{k,ij} + \tau^{-1}\tau_0\sum_{(i,j)\in P_q} m_{q,ij}}\right\} \\
&- A_n^{-1}\sum_{k=0}^{p}\sum_{(i,j)\in P_k}\log|\alpha_{k,ij}(d,\tau)|.
\end{aligned}$$

Under Assumption 1. (4), as n tends to be large,

$$\sup_{k=0,\dots,p,(i,k)\in P_k} \sup_{(d,\tau)\in D\times\Theta} \left|tr(\alpha_{k,ij}(d,\tau) - I_{m_{k,ij}})\right| \to 0.$$

Hence we obtain

$$\sup_{(d,\tau)\in D\times\Theta}|R_n(d,\tau)| \le K \sup_{k=0,\dots,p,(i,k)\in P_k} \sup_{(d,\tau)\in D\times\Theta} \left|tr(\alpha_{k,ij}(d,\tau) - I_{m_{k,ij}})\right| \to 0.$$

Applying the argument of Proposition A.1 in Hurvich et al. (2005), summation by parts and Assumtion 1. (4), we obtain that

$$\sup_{(d,\tau)\in D\times\Theta} \left| \log\left(1 + \sum_{k=0}^{p} \sum_{(i,j)\in P_k} \gamma_{k,ij}(d,\tau) + \sum_{(i,j)\in P_q} \gamma_{q,ij}(d,\tau)\right)\right| \to 0,$$

in probability, by which we have the consistency of $\hat{d}$.

7.2.2. *Consistency of* $\hat{\tau}$. Define

$$\begin{aligned}
J_n(x) =& \log A_n^{-1}\left(\sum_{k=0}^{p}\sum_{(i,j)\in P_k} m_{k,ij} + x\sum_{(i,j)\in P_q} m_{q,ij}\right) \\
&- A_n^{-1}\sum_{(i,j)\in P_q} m_{q,ij}\log x.
\end{aligned}$$

The function $J_n(x)$ is twice differentialble, $J_n(1) = 0$, $J_n'(1) = 0$ and $J_n''(x)$ is bounded away from 0 for $x \in [0,2]$. Thus there exists a constant c such that for $\tau \in \Theta$,

$$J_n(\tau^{-1}\tau_0) \ge c\left(\tau^{-1}\tau_0 - 1\right)^2.$$

Hence, defining $S_n(\tau) = l_0(\hat{d},\tau) - l_0(d_0,\tau_0) - J_n(\tau^{-1}\tau_0)$, we obtain

$$\begin{aligned}
0 \le (\hat{\tau}^{-1}\tau_0 - 1)^2 1_\Theta(\hat{\tau}) &\le c^{-1}J_n(\hat{\tau}^{-1}\tau_0) = c^{-1}l_0(\hat{d},\tau) - c^{-1}l_0(d_0,\tau_0) - c^{-1}S_n(\hat{\tau}) \\
&\le c^{-1}\sup_{(d,\tau)\in D\times\Theta}|l_0(d,\tau) - l_0(d_0,\tau_0)| + c^{-1}\sup_{\tau\in\Theta}|S_n(\tau)|,
\end{aligned}$$

and

$$S_n(\tau) = \log\left\{1 + \frac{\sum_{k=0}^{p}\sum_{(i,j)\in P_k}(k+1)^{\hat{d}-d_0} tr\left(\alpha_{k,ij}(\hat{d},\tau)\right) - m_{k,ij}}{\sum_{k=0}^{p}\sum_{(i,j)\in P_k} m_{k,ij} + \tau^{-1}\tau_0\sum_{(i,j)\in P_q} m_{q,ij}}\right\}$$
$$- A_n^{-1}\sum_{k=0}^{p}\sum_{(i,j)\in P_k}\log\left|(k+1)^{\hat{d}-d_0}\alpha_{k,ij}(\hat{d},\tau)\right|.$$

By the consistency of $\hat{d}$ and Assumption 1. (4), we obtain

$$\sup_{\tau\in\Theta}|S_n(\tau)| \to 0,$$

in probability, which proves the consistency of $\hat{\tau}$.

7.2.3. *Consistency of $\hat{C}$.* Define

$$\begin{aligned} e_{k,ij} &= C_0^{-1}k^{d_0}H_{k,ij}^{-1}(d_0,\tau_0)W_{k,ij} - I_{m_{k,ij}}, \\ e_{q,ij} &= C_0^{-1}\tau_0^{-1}B_{q,ij}^{-1}W_{q,ij} - I_{m_{q,ij}}. \end{aligned}$$

Then

$$\begin{aligned} &\hat{C} - C_0 = \\ &C_0A_n^{-1}\left[\sum_{k=0}^{p}\sum_{(i,j)\in P_k} k^{\hat{d}-d_0} tr\left\{\alpha_{k,ij}(\hat{d},\hat{\tau})e_{k,ij}\right\} + \sum_{(i,j)\in P_q}\hat{\tau}^{-1}\tau_0 tr\left(e_{q,ij}\right)\right] \\ &+C_0A_n^{-1}\left[\sum_{k=0}^{p}\sum_{(i,j)\in P_k} k^{\hat{d}-d_0} tr\left\{\alpha_{k,ij}(\hat{d},\hat{\tau}) - I_{m_{k,ij}}\right\} + \sum_{(i,j)\in P_q} m_{q,ij}\left(\hat{\tau}^{-1}\tau_0 - 1\right)\right] \\ &\to 0 \end{aligned}$$

in probability.

References

[1] Bai, Y., Song, P. and Raghunathan, T. E. (2012) Joint composite estimating functions in spatio temporal models. *J. R. Statist. Soc.* B, **74**, 799-824.

[2] Banerjee, S., Carlin, B. and Gelfand A. (2004) *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC.

[3] Banerjee, S., Gelfand A., Finley, A. O. and Sang, H. (2008) Gaussian predictive process models for large spatial data sets. *J. R. Statist. Soc.* B, **70**, 825–848.

[4] Brockwell, P. J. and Davis, R. A. (1991) *Time Series: Theory and Methods*. Springer-Verlag, 2nd ed.

[5] Cressie, N. and Johannesson, G. (2008) Fixed rank kriging for very large spatial data sets. *J. R. Statist. Soc. B* , 70, 209-226.

[6] Donoho, D. L. and Johnstone, I. M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81, 425-455.

[7] Hurvich, C., Moulines, E. and Soulier, P. (2005) Estimating long memory in volatility. *Econometrika*, 73, 1283-1328.

[8] Kaufman, C. G., Schervish, M. J. and Nychka, W. (2008) Covariance tapering for likelihood-based estimation in large spatial data sets. *J. Am. Statist. Assoc.*, **103**, 1545–55.

[9] Ogden, R. T. (1997) *Essential Wavelets for Statistical Applications and Data Analysis*. BirkHäuser, Boston.

[10] Paciorek, C. J. (2007) Bayesian smoothing with Gaussian processes using Fourier basis functions in the spectral GP package. *J. Statist. Software*, 19, 1-28.

[11] Robinson, P. M. (1995) Gaussian semiparametric estimation of long range dependence. *Annals of Statistics*, 24, 1630-1661.

[12] Shi, T. and Cressie, N. (2007) Global statistical analysis of MISR aerosol data: a massive data product from NASA's Terra satellite. *Environmetrics*, 18, 665-680.
[13] Stein, M. L. (1999) *Interpolation of Spatial Data*. New York: Springer-Verlag.
[14] Stroud, J., Mueller and Sanso, B. (2001) Dynamic models for spatial-temporal data. *J. Royal Statist. Soc. B.*, 63, 673-689.
[15] Wikle, C. K. (2010) Low-rank representations for spatial processes. *Handbook of Spatial Statistics.* (Edited by A. E. Gelfand, P. J. Diggle, M. Fuentes and P. Guttorp), 107-118. Chapman & Hall/CRC.
[16] Zimmerman, D. L. and Stein, M. (2010) Classical geostatistical methods. *Handbook of Spatial Statistics.* (Edited by A. E. Gelfand, P. J. Diggle, M. Fuentes and P. Guttorp), 29-44. Chapman & Hall/CRC.

Graduate School of Economics and Management, Tohoku University, 27-1 Kawauchi, Aoba ward, Sendai 980-8576, Japan
E-mail address: matsuda@econ.tohoku.ac.jp